DETAILED REPORT

SECOND REVIEW

COVID-19 DATA ANALYSIS AND FORECASTING

GUIDE SIGNATURE

SUBMITTED BY

REAHAAN SHERIFF I

2019202045

MCA – R 3YRS

COVID-19 DATA ANALYSIS AND FORECASTING

A PROJECT REPORT

Submitted by

REAHAAN SHERIFF I – 2019202045

A report of the project submitted to the Faculty of

INFORMATION SCIENCE AND TECHNOLOGY

in partial fulfillment

for the award of the degree

of

MASTER OF COMPUTER APPLICATION

DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY,

COLLEGE OF ENGINEERING, GUINDY

ANNA UNIVERSITY CHENNAI 600 025

MAY, 2022

ANNA UNIVERSITY

CHENNAI - 600 025

BONAFIDE CERTIFICATE

Certified that this project report titled COVID-19 DATA ANLYSIS AND FORECASTING is the bonafide work of REAHAAN SHERIFF I who carried out project work under my supervision. Certified further that to the best of my knowledge and belief, the work reported herein does not form part of any other thesis or dissertation based on which a degree or an award was conferred on an earlier occasion on this or any other candidate.

PLACE: CHENNAI

DATE: 30-05-2022

PROJECT GUIDE

Ms. P.S APIRAJITHA

TEACHING FELLOW

DEPARTMENT OF IST

ANNA UNIVERSITY

CHENNAI 600025

DR.S.SRIDHAR

HEAD OF THE DEPARTMENT

DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY

COLLEGE OF ENGINEERING, GUINDY

ANNA UNIVERSITY

CHENNAI 600025

ABSTRACT

COVID-19 has sparked a worldwide pandemic, with the number of infected cases and deaths rising on a regular basis. Along with recent advances in soft computing technology, researchers are now actively developing and enhancing different mathematical and machine-learning algorithms to forecast the future trend of this pandemic. Thus, if we can accurately forecast the trend of cases globally, the spread of the pandemic can be controlled. In this project, a LSTM model, Prophet Model and ARIMA model will be used on a time-series dataset to forecast the cases of COVID-19 in future. Machine learning (ML) is a category of an algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. Long Short Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long-term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give an efficient performance. Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well. An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends. A statistical model is autoregressive if it predicts future values based on past values.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

Machine learning (ML) is a category of an algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. The basic premise of machine learning is to build algorithms that can receive input data and use statistical analysis to predict an output while updating outputs as new data becomes available. Long Short Term Memory is a kind of recurrent neural network. In RNN output from the last step is fed as input in the current step. It tackled the problem of long-term dependencies of RNN in which the RNN cannot predict the word stored in the long-term memory but can give more accurate predictions from the recent information. As the gap length increases RNN does not give an efficient performance. Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well. An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends. A statistical model is autoregressive if it predicts future values based on past values.

1.1 MOTIVATION AND OBJECTIVE:

To predict the cases and peak case date for COVID-19 in India as accurately as possible. We will be using LSTM, Prophet and ARIMA machine learning models.

1.2 IMPLEMENTATION PLATFORM / FRAMEWORK:

Google COLAB used to train the model using machine learning algorithms.

Libraries used,

1. **Pandas** - It is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language
2. **Numpy** - It offers comprehensive mathematical functions, random number generators, linear algebra routines, Fourier transforms, and more. Interoperable.

3. **Seaborn** - It is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical data.
4. **Tensorflow** - It is an end-to-end open source platform for machine learning. It has a comprehensive, flexible ecosystem of tools, libraries and community resources that lets researchers push the state-of-the-art in ML and developers easily build and deploy ML powered applications.
5. **Matplotlib** - It is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

To interact with the machine learning model a front end user interface is build using HTML, CSS, BOOTSTRAP and integrated the machine learning model with backend server DJANGO.

1. **HTML** - The HyperText Markup Language or HTML is the standard markup language for documents designed to be displayed in a web browser. It can be assisted by technologies such as Cascading Style Sheets (CSS) and scripting languages such as JavaScript.
2. **CSS** - Cascading Style Sheets (CSS) is a stylesheet language used to describe the presentation of a document written in HTML or XML (including XML dialects such as SVG, MathML or XHTML). CSS describes how elements should be rendered on screen, on paper, in speech, or on other media.
3. **BOOTSTRAP** - Bootstrap is a potent front-end framework used to create modern websites and web apps. It's open-source and free to use, yet features numerous HTML and CSS templates for UI interface elements such as buttons and forms. Bootstrap also supports JavaScript extensions.
4. **DJANGO** - Django is a high-level Python web framework that encourages rapid development and clean, pragmatic design. Built by experienced developers, it takes care of much of the hassle of web development, so you can focus on writing your app without needing to reinvent the wheel. It’s free and open source.

VS code - Visual Studio Code is a lightweight but powerful source code editor which runs on your desktop and is available for Windows, macOS and Linux. It comes with built-in support for JavaScript, TypeScript and Node.js and has a rich ecosystem of extensions for other languages (such as C++, C#, Java, Python, PHP, Go) and runtimes (such as .NET and Unity).

CHAPTER 2

LITERATURE REVIEW

2.1 Coronavirus disease (COVID-19) cases analysis using machine-learning applications

Today world thinks about coronavirus disease that which means all even this pandemic disease is not unique. The purpose of this study is to detect the role of machine-learning applications and algorithms in investigating and various purposes that deals with COVID-19. Review of the studies that had been published during 2020 and were related to this topic by seeking in Science Direct, Springer, Hindawi, and MDPI using COVID-19, machine learning, supervised learning, and unsupervised learning as keywords. The total articles obtained were 16,306 overall but after limitation; only 14 researches of these articles were included in this study. Our findings show that machine learning can produce an important role in COVID-19 investigations, prediction, and discrimination. In conclusion, machine learning can be involved in the health provider programs and plans to assess and triage the COVID-19 cases. Supervised learning showed better results than other Unsupervised learning algorithms by having 92.9% testing accuracy. In the future recurrent supervised learning can be utilized for superior accuracy.

2.2 Optimizing LSTM for time series prediction in Indian stock market

Long Short Term Memory (LSTM) is among the most popular deep learning models used today. It is also being applied to time series prediction which is a particularly hard problem to solve due to the presence of long term trend, seasonal and cyclical fluctuations and random noise. The performance of LSTM is highly dependent on choice of several hyper-parameters which need to be chosen very carefully, in order to get good results. Being a relatively new model, there are no established guidelines for configuring LSTM. In this paper this research gap was addressed. A dataset was created from the Indian stock market and an LSTM model was developed for it. It was then optimized by comparing stateless and stateful models and by tuning for the number of hidden layers.

2.3: A Comparative Study:Time-Series Analysis Methods for Predicting COVID-19 Case Trend

Since 2019, COVID-19, as a new acute respiratory disease, has struck thewhole world, causing millions of death and threatening the economy, politics, and civilization. Therefore, an accurate prediction of the future spread of COVID-19 becomes crucial in such a situation. In this comparative study, four different time-series analysis models, namely the ARIMA model, the Prophet model, the Long Short-Term Memory (LSTM) model, and the Transformer model, are investigated to determine which has the best performance when predicting the future case trends of COVID-19 in six countries. After obtaining the publicly available COVID-19 case data from Johns Hopkins University Center for Systems Science and Engineering database, we conduct repetitive experiments which exploit the data to predict future trends for all models. The performance is then evaluated by mean squared error (MSE) and mean absolute error (MAE) metrics. The results show that overall the LSTM model has the best performance for all countries that it can achieve extremely low MSE and MAE. The Transformer model has the second-best performance with highly satisfactory results in some countries, and the other models have poorer performance. This project highlights the high accuracy of the LSTM model, which can be used to predict the spread of COVID-19 so that countries can be better prepared and aware when controlling the spread.

CHAPTER 3

SYSTEM DESIGN

This chapter consists of the system design of the project with the overall architecture and the description of the modules used in the project

3.1 OVERALL ARCHITECTURE:

The overall system architecture of the proposed system is shown in Figure 3.1. For analysis the time series dataset will be taken and it will be pre-processed. That pre-processed data will be splitted into training and testing. The trained dataset will be applied to the LSTM model, Prophet model and ARIMA model these models will be used for evaluating the test dataset. After the results the model will be used for future forecasting to forecast the upcoming cases and the evaluation of those results will be done in RMSE, MSE and MAE metrics.

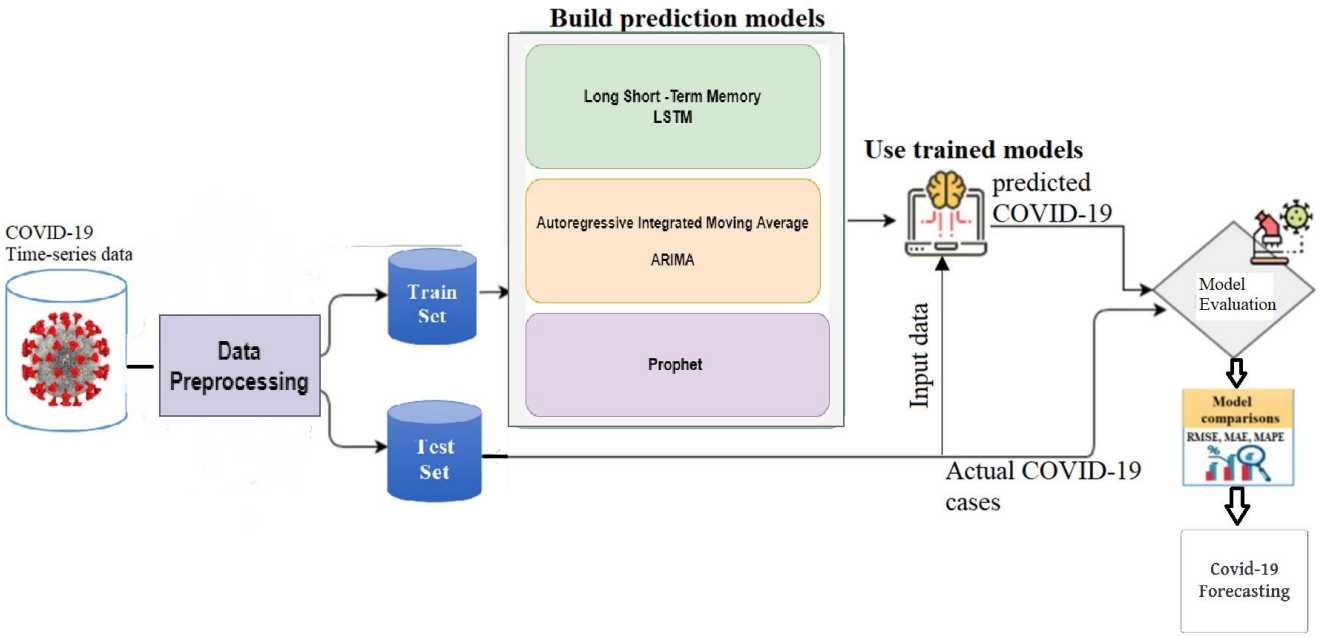

Figure 3.1: Architecture Diagram

3.2 MODULES DESCRIPTION:

3.2.1 Data pre-processing

Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model.

3.2.2 Building model

LSTM:
I will built the model with the help of LSTM. The model has an input layer followed by three LSTM layers. The LSTM layers contain Dropout as 0.2 to prevent overfitting in the model. The output layer consists of a dense layer with 1 neuron with activation as relu. We will predict the number of Corona cases, so our output will be a positive number $(0, \infty)$.

Prophet:
Prophet is a procedure for forecasting time series data based on an additive model where non-linear trends are fit with yearly, weekly, and daily seasonality, plus holiday effects. It works best with time series that have strong seasonal effects and several seasons of historical data. Prophet is robust to missing data and shifts in the trend, and typically handles outliers well.

ARIMA:
An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends. A statistical model is autoregressive if it predicts future values based on past values

3.2.3 Training model

To train the model, we will take out training data (80%) and used 20% of it as validation data. To higher the learning rate of our model we will use time series generator in the model. Training the model with n epochs.

3.2.4 Model evaluation

It estimates how well (or how bad) the model is, in terms of its ability in mapping the relationship between X (a feature, or independent variable, or predictor variable) and Y (the target, or dependent variable, or response variable).

3.2.5 Visualization and forecasting

In order to see the prediction and accuracy, first, we predicted the output of our x_test data. This was the output that we got from the test data. To accurately plot the values, we needed to bring our prediction and y_test data back to the original bounds of the data. In the end, we plotted a graph between the actual COVID-19 cases compared to our predicted COVID-19 cases to see the overall accuracy of our model

CHAPTER 4

MODULE-BASED OUTPUT SCREENSHOTS

Home page:

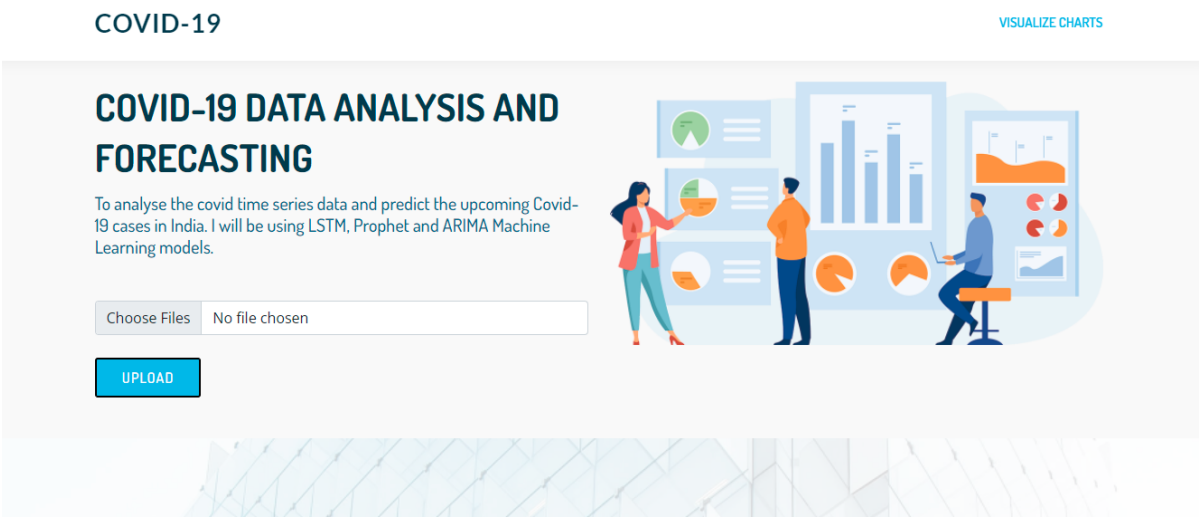

Figure 4.1: Home screen

This image displays the required field to upload the datasets which is used in the project

Files:

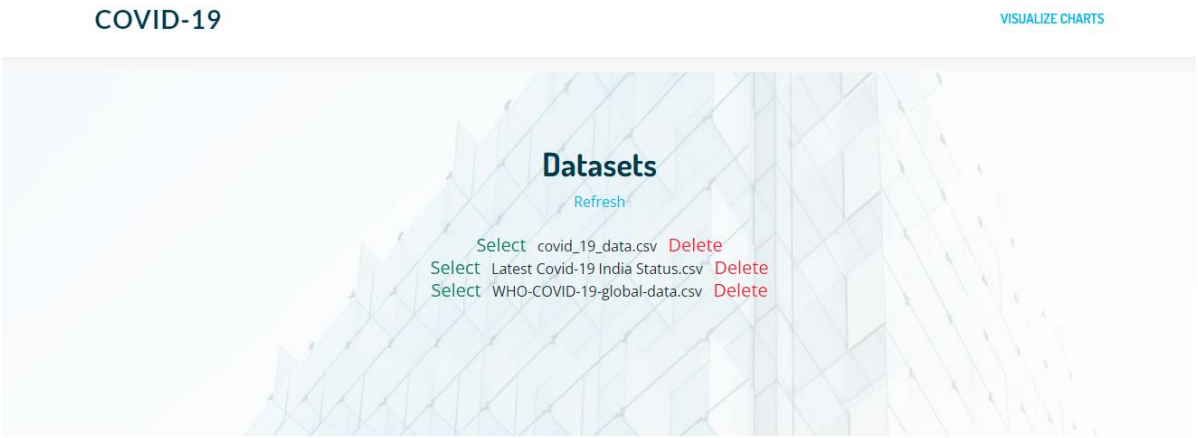

Figure 4.2: Datasets

This image displays the required datasets that will be used in this project

Machine Learning models:

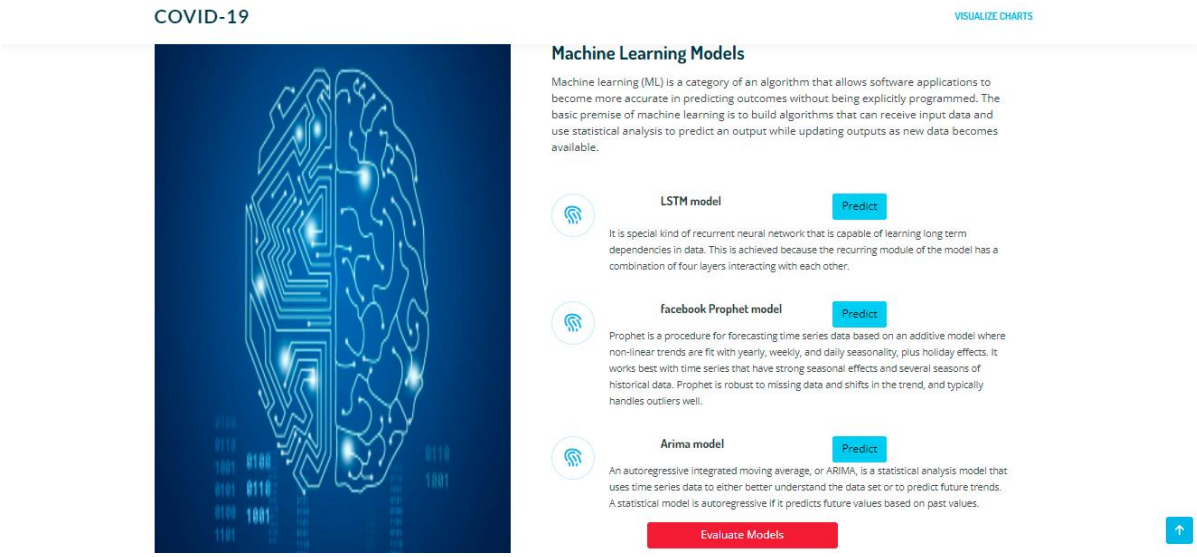

Figure 4.3: Models

This image displays the machine learning models that will be used in this project

Exploratory data analysis:

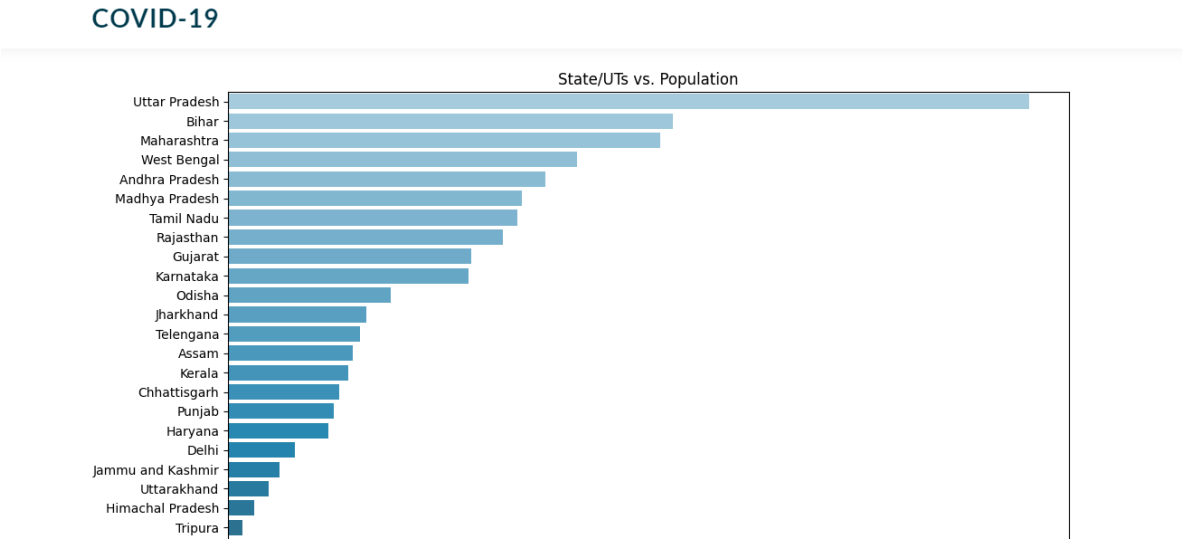

Figure 4.2.1: EDA

This image displays the total population analysis that done on the dataset.

COVID-19

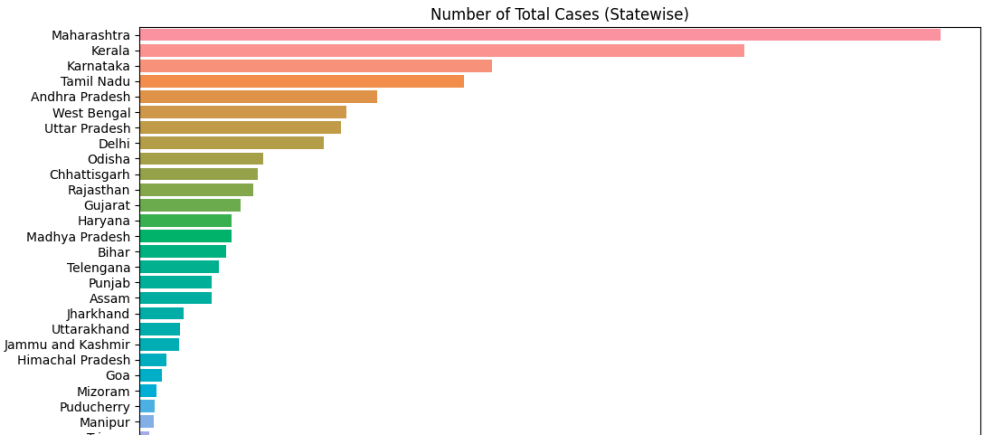

Figure 4.2.2: EDA

This image displays the no of cases analysis that done on the dataset.

COVID-19

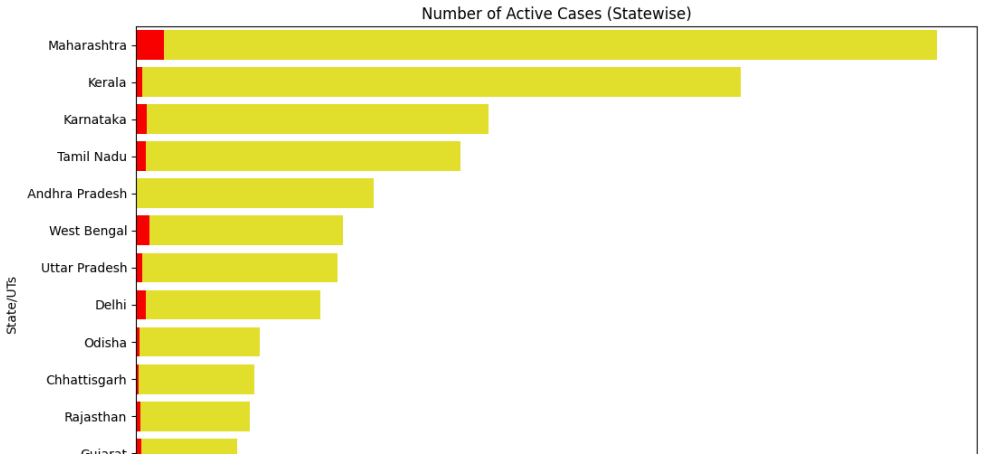

Figure 4.2.3: EDA

This image displays the no of cases analysis that done on the dataset

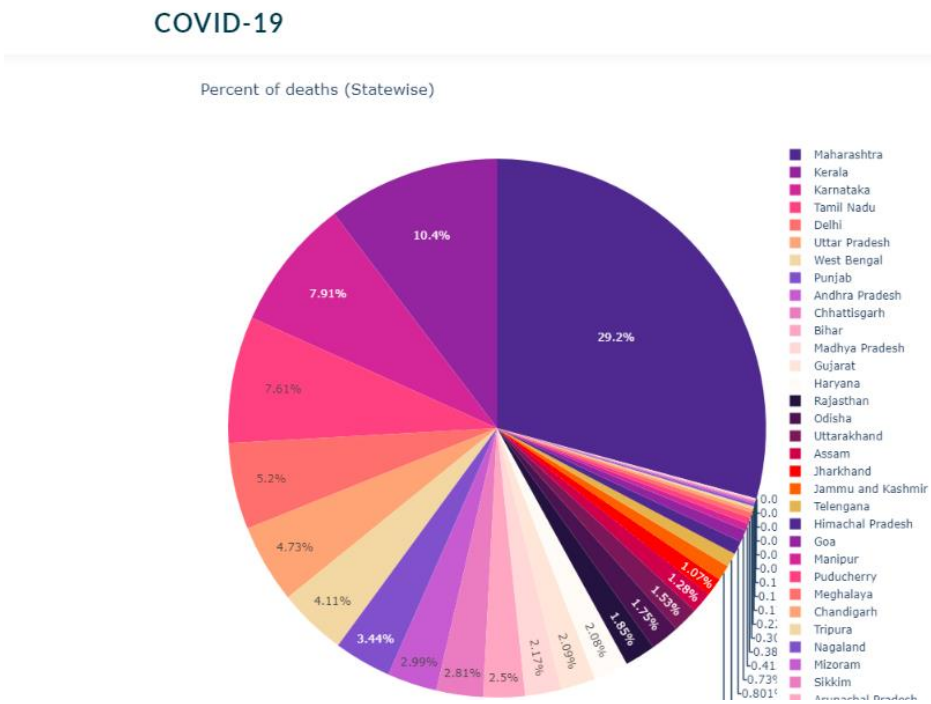

Figure 4.2.4: EDA

This image displays the no of death analysis that done on the dataset

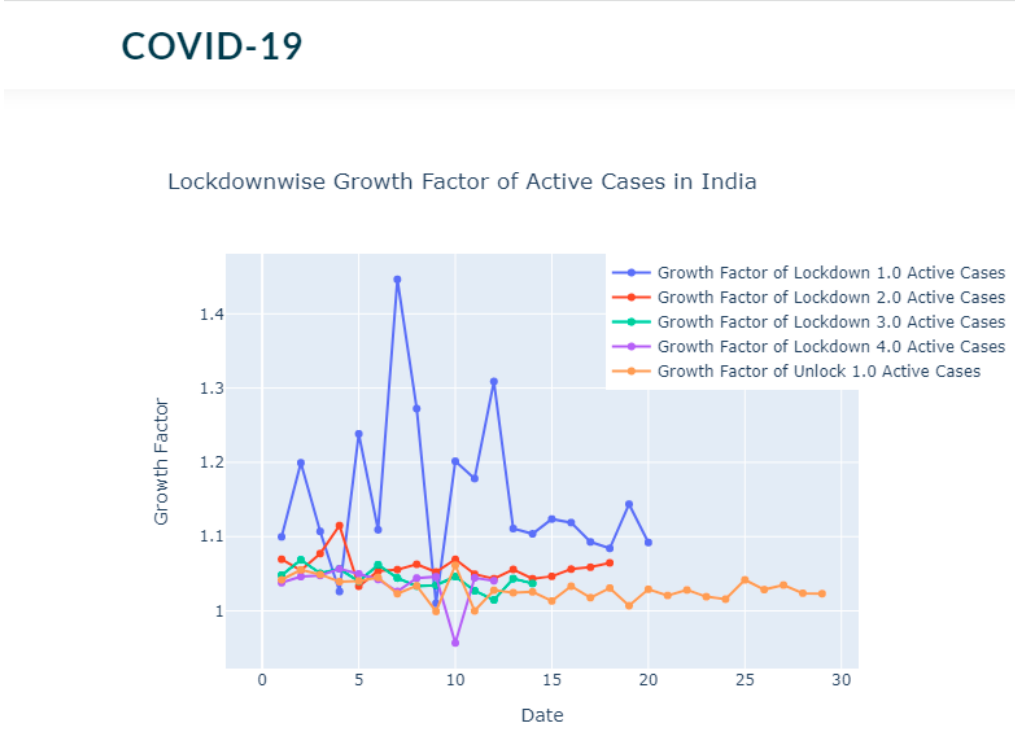

Figure 4.2.5: EDA

This image displays the analysis of cases on lockdown period.

LSTM model prediction:

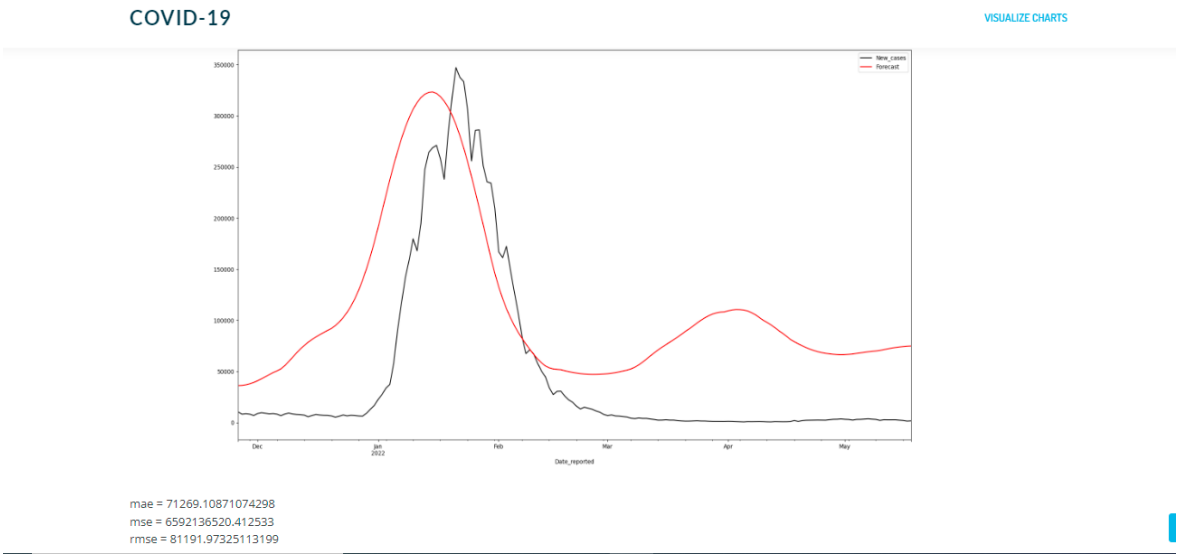

Figure 4.3: LSTM

This image displays the actual and predicted cases of covid 19 using LSTM model.

Prophet model:

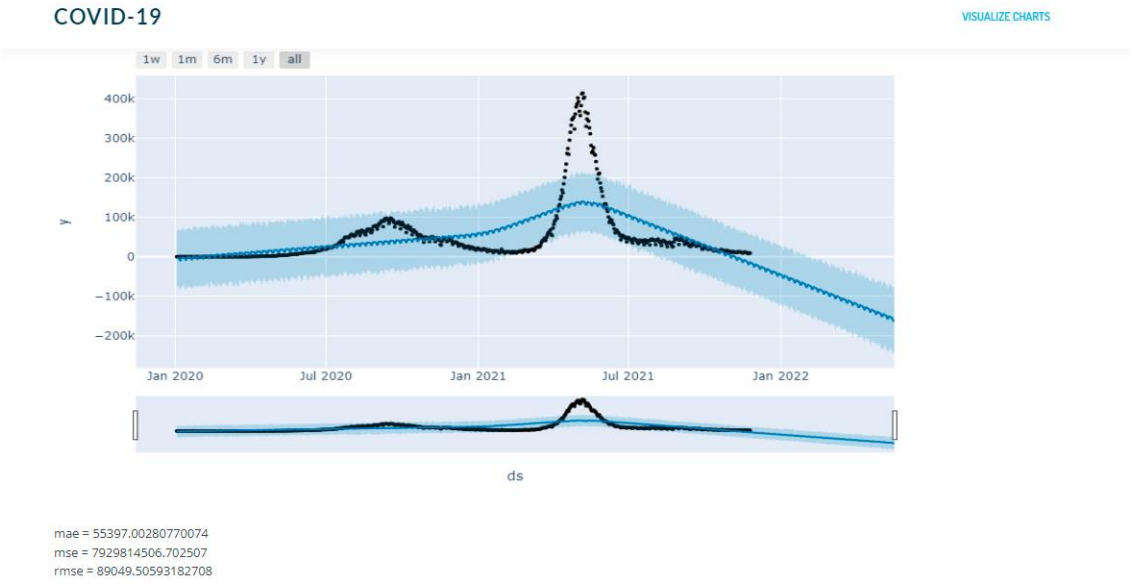

Figure 4.4: Prophet

This image displays the actual and predicted cases of covid 19 using Prophet model.

ARIMA model:

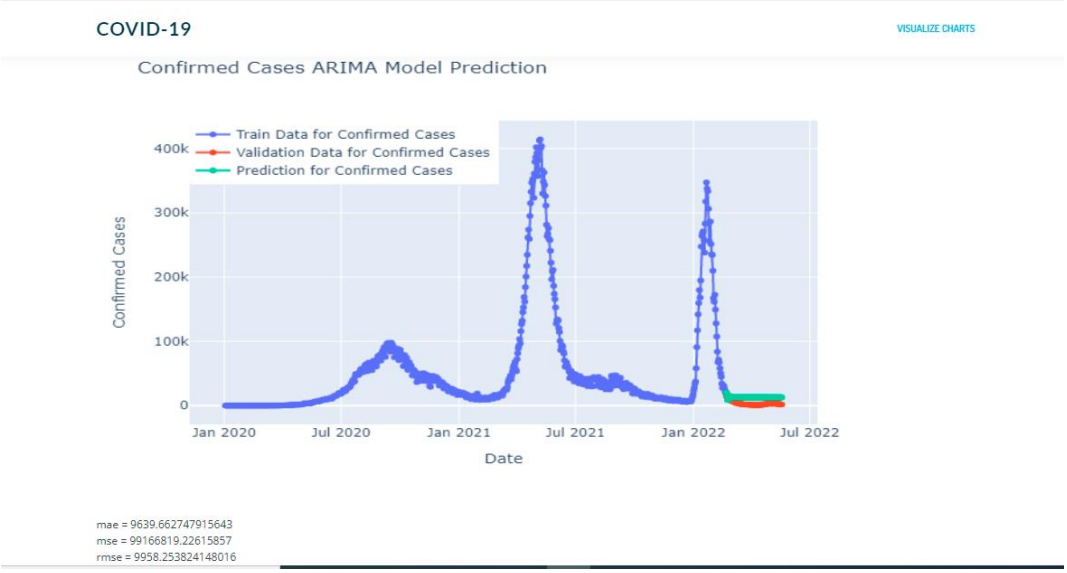

Figure 4.5: ARIMA

This image displays the actual and predicted cases of covid 19 using ARIMA model.

Evaluation metrics:

	Model Name	Mean Squared Error	Mean Absolute Error	Root Mean Squared Error
2	ARIMA Model	99166819.226159	9639.662748	9958.253824
0	LSTM Model	6592136520.412533	71269.108711	81191.973251
1	Prophet Model	7929814506.702507	55397.002808	89049.505932

Figure 4.3: Metrics

This image displays the evaluation metrics of the performed machine learning model.

REFERENCES:

1. Wang J., Liu Y., Wei Y., Xia J., Yu T., Zhang X., Zhang L. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. Lancet. 2020;395(10223):507–513.

2. JA Backer, D Klinkenberg and J. Wallinga, Incubation period of 2019 novel coronavirus (2019-nCoV) infections among travellers from Wuhan, China, 20–28 January 2020.

3. Tolga Ergen and Suleyman Serdar Kozat, "Efficient online learning algorithms based on LSTM neural networks", IEEE transactions on neural networks and learning systems, vol. 29, no. 8, pp. 3772-3783, 2017.

4. H. Li, S.-M. Liu, X.-H. Yu, S.-L. Tang and C.-K. Tang, "Coronavirus disease 2019 (COVID-19): current status and future perspectives", International Journal of Antimicrobial Agents, pp. 105951, 2020.

5. Ameer Sardar, Kwekha Rashid, Heamn N. Abduljabbar & Bilal Alhayani, Applied Nanoscience (2021) “Coronavirus disease (COVID-19) cases analysis using machine-learning applications”